Recognition of Reverberant Speech by Missing Data Imputation and NMF Feature Enhancement

Heikki Kallasjoki*, Jort F. Gemmeke, Kalle J. Palomäki, Amy V. Beeston, Guy J. Brown

Department of Signal Processing and Acoustics
Aalto University, School of Electrical Engineering
heikki.kallasjoki@aalto.fi
http://research.spa.aalto.fi/speech/robust/kallasjoki-reverb14/

May 10, 2014
Outline

Introduction

Methods
  Missing data imputation
  NMF-based feature enhancement
  Further processing

Results

Conclusions
Introduction

- Two lines of investigation:
  - Missing data methods for dereverberation
  - Extending NMF-based feature enhancement
- Both turn out to be beneficial for reverberant speech (even with multi-condition training, CMLLR adaptation)
Outline

Introduction

Methods
   Missing data imputation
   NMF-based feature enhancement
   Further processing

Results

Conclusions
Missing Data Framework

- Essential idea: focus on spectro-temporal regions dominated by the speech signal
- Estimate reliability (soft or hard decision)
- Use the estimates to improve speech recognition (e.g. by marginalization, imputation...)
- Can make minimal assumptions about the distortion
- In this work: feature imputation with binary masks
Mask Estimation

$m_R$

$m_{GMM}$

$m_{LP}$

$m_{SVM}$
Mask Estimation: $m_R$

- Based on mel-spectral features compressed to $x^{0.3}$
- Band-pass modulation filter, 1.5...8.2 Hz
- Followed by an AGC and normalization
- Threshold based on “blurredness” metric: ratio of channel mean and channel max
Mask Estimation: $m_R$, illustrated
Mask Estimation: \( m_{LP} \)

- Based on normalized \( x^{0.3} \) mel-spectral features
- Low-pass modulation filter with cutoff at 10 Hz
- Means of each contiguous region where \( y' < 0 \)
Mask Estimation: $m_{\text{GMM}}$ & $m_{\text{SVM}}$

- Oracle mask:
  - threshold difference between clean and reverberant
- Features: spectra, gradient, “blurredness”, $m_R$, $m_{\text{LP}}$
- Train a (GMM or SVM) classifier for each channel
Bounded Conditional Mean Imputation

Conditional Mean Imputation

- Model distribution of clean speech $\mathbf{x}$ with a GMM
- Estimate missing $\mathbf{x}_u$ by conditioning on reliable $\mathbf{x}_r$:

$$\hat{\mathbf{x}}_u = \int_{\mathbf{x}_u} \mathbf{x}_u p(\mathbf{x}_u | \mathbf{x}_r)$$

Bounded Conditional Mean Imputation

- Use observation as upper bound: $\hat{\mathbf{x}}_u < \mathbf{x}_u^{obs}$
- In this work:
  truncated $p(\mathbf{x}_u | \mathbf{x}_r)$ approximated with a parametric model
Outline

Introduction

Methods
- Missing data imputation
- NMF-based feature enhancement
- Further processing

Results

Conclusions
NMF Signal Model

\[ 0.50 + 0.25 + 0.15 + \cdots = \]
Using NMF for Speech Feature Enhancement

Example: source separation for noisy speech

- Fixed dictionary of clean speech and noise samples (also called exemplars)
- After solving coefficients, reconstruct clean speech only
- A lot of flexibility here
Using NMF for Speech Feature Enhancement

Example: source separation for noisy speech

- Fixed dictionary of clean speech and noise samples (also called *exemplars*)
- After solving coefficients, reconstruct clean speech only
- A lot of flexibility here

What about reverberation?

- Source separation approach not directly applicable
Accounting for Reverberation

\[ Y \approx S A \]

\[ TC \times W \]
stacked observation

\[ TC \times N \]
dictionary matrix

\[ N \times W \]
activation matrix

\[ \text{Exercise: Modeling with a reverberated dictionary} \]

\[ \text{Exercise: Reverberating the NMF approximation} \]
Accounting for Reverberation

\[ Y \approx R S A \]

- \( T_r C \times W \): stacked observation
- \( T_r C \times TC \): filter matrix
- \( TC \times N \): dictionary matrix
- \( N \times W \): activation matrix

- (RS) A: modeling with a reverberated dictionary
- R (SA): reverberating the NMF approximation
The Filter Matrix $R$

$$R = \begin{pmatrix}
0 & 0 & 0 \\
0 & r_{1,2} & 0 & \ldots \\
0 & 0 & r_{1,3} & \\
\vdots & \vdots & \ddots & \\
0 & 0 & 0 & r_{1,1} \\
0 & r_{2,2} & 0 & \ldots & 0 & r_{1,2} & 0 & \ldots \\
0 & 0 & r_{2,3} & 0 & 0 & r_{1,3} \\
\vdots & \vdots & \ddots & \ddots & \ddots
\end{pmatrix}$$

$$= T_{rC}$$
Issues

- Does not want to converge to a useful solution

- Sliding-window approach not so suitable for reverberation
Issues

- Does not want to converge to a useful solution
  - Initialization with missing-data imputation
  - Tuning of iteration scheme
  - Activation matrix filtering
- Sliding-window approach not so suitable for reverberation
  - Sum overlapping windows in multiplicative updates
  - (Or do convolutive NMF)
The Case for Convolutional NMF
The Case for Convolutional NMF
NMF Feature Enhancement Process

1. Estimate $\tilde{X}$ using BCMI
2. Iteratively update $A$ in $\tilde{X} \approx RSA$ with identity $R$
3. Filter $A$ to suppress consecutive nonzero activations
4. Initialize $R$ to contain filter $\frac{1}{T_f} [1 \ldots 1]$ on all channels
5. Iteratively update $R$ in $Y \approx RSA$ with fixed $A$
   (under constraints $r_{t+1,b} < r_{t,b}$, $\sum_{t,b} r_{t,b} = C$)
6. Iteratively update $A$ in $Y \approx RSA$ with fixed $R$
   ▶ Then use $\hat{X} = SA$ and $\hat{Y} = RSA$ for feature enhancement,
   with a per-frame Wiener filter in the mel-spectral domain
Outline

Introduction

Methods
- Missing data imputation
- NMF-based feature enhancement
- Further processing

Results

Conclusions
Further Processing

Channel Normalization
- Mean of the $\frac{1}{L}$ largest-valued samples on each channel
- Reduces mismatch between NMF dictionary and test data

Beamforming
- Simple delay-sum beamformer
- TDOA estimation with PHAT-weighted cross-correlation
Outline

Introduction

Methods
  Missing data imputation
  NMF-based feature enhancement
  Further processing

Results

Conclusions
Setup

- REVERB Challenge HTK recognizer
- Four sets of acoustic models:
  - **Clean** WSJCAM0 clean speech training set
  - **MC** REVERB Challenge multi-condition training set
  - **MC+ad.** . . . with CMLLR adaptation over a test condition
  - **8-ch.** . . . on audio preprocessed with the PHAT-DS beamformer
Results for Mask Estimation Methods

- Development set, clean speech acoustic models

<table>
<thead>
<tr>
<th>Method</th>
<th>SimData</th>
<th>RealData</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>51.81</td>
<td>88.51</td>
</tr>
<tr>
<td>BCMI mask (m_R)</td>
<td>40.07</td>
<td>67.88</td>
</tr>
<tr>
<td>BCMI mask (m_{LP})</td>
<td>48.01</td>
<td>73.06</td>
</tr>
<tr>
<td>BCMI mask (m_{GMM})</td>
<td>39.94</td>
<td>70.87</td>
</tr>
<tr>
<td>BCMI mask (m_{SVM})</td>
<td>40.78</td>
<td>74.14</td>
</tr>
<tr>
<td>NMF (with (m_R))</td>
<td>28.26</td>
<td>58.84</td>
</tr>
</tbody>
</table>
### Results for Mask Estimation Methods

- Development set, clean speech acoustic models

<table>
<thead>
<tr>
<th>Method</th>
<th>SimData</th>
<th>RealData</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>51.81</td>
<td>88.51</td>
</tr>
<tr>
<td>BCMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mask $m_R$</td>
<td>40.07</td>
<td>67.88</td>
</tr>
<tr>
<td>mask $m_{LP}$</td>
<td>48.01</td>
<td>73.06</td>
</tr>
<tr>
<td>mask $m_{GMM}$</td>
<td>39.94</td>
<td>70.87</td>
</tr>
<tr>
<td>mask $m_{SVM}$</td>
<td>40.78</td>
<td>74.14</td>
</tr>
<tr>
<td>NMF (with $m_R$)</td>
<td>28.26</td>
<td>58.84</td>
</tr>
</tbody>
</table>
## Results for Feature Enhancement

<table>
<thead>
<tr>
<th>Model</th>
<th>FE</th>
<th>SimData</th>
<th>RealData</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Baseline</td>
<td>51.82</td>
<td>89.04</td>
</tr>
<tr>
<td></td>
<td>BCMI</td>
<td>39.14</td>
<td>71.67</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td><strong>29.74</strong></td>
<td><strong>59.13</strong></td>
</tr>
<tr>
<td>MC</td>
<td>Baseline</td>
<td>29.60</td>
<td>56.58</td>
</tr>
<tr>
<td></td>
<td>BCMI</td>
<td>27.25</td>
<td>51.31</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td><strong>24.11</strong></td>
<td><strong>47.06</strong></td>
</tr>
<tr>
<td>MC+ad.</td>
<td>Baseline</td>
<td>25.37</td>
<td>48.88</td>
</tr>
<tr>
<td></td>
<td>BCMI</td>
<td>24.58</td>
<td>46.05</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td><strong>21.91</strong></td>
<td><strong>41.41</strong></td>
</tr>
<tr>
<td>8-ch.</td>
<td>Baseline</td>
<td>19.76</td>
<td>40.21</td>
</tr>
<tr>
<td></td>
<td>BCMI</td>
<td>19.40</td>
<td>38.28</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td><strong>17.80</strong></td>
<td><strong>34.79</strong></td>
</tr>
</tbody>
</table>
Results for Feature Enhancement

<table>
<thead>
<tr>
<th>Model</th>
<th>FE</th>
<th>SimData</th>
<th>RealData</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Baseline</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Clean</td>
<td>BCMI</td>
<td>–24.5%</td>
<td>–19.5%</td>
</tr>
<tr>
<td>Clean</td>
<td>NMF</td>
<td>−42.6%</td>
<td>−33.6%</td>
</tr>
<tr>
<td>MC</td>
<td>Baseline</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MC</td>
<td>BCMI</td>
<td>−7.9%</td>
<td>−9.3%</td>
</tr>
<tr>
<td>MC</td>
<td>NMF</td>
<td>−18.5%</td>
<td>−16.8%</td>
</tr>
<tr>
<td>MC+ad.</td>
<td>Baseline</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MC+ad.</td>
<td>BCMI</td>
<td>−3.1%</td>
<td>−5.8%</td>
</tr>
<tr>
<td>MC+ad.</td>
<td>NMF</td>
<td>−13.6%</td>
<td>−15.3%</td>
</tr>
<tr>
<td>8-ch.</td>
<td>Baseline</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>8-ch.</td>
<td>BCMI</td>
<td>−1.8%</td>
<td>−4.8%</td>
</tr>
<tr>
<td>8-ch.</td>
<td>NMF</td>
<td>−9.9%</td>
<td>−13.5%</td>
</tr>
</tbody>
</table>
Outline

Introduction

Methods
  Missing data imputation
  NMF-based feature enhancement
  Further processing

Results

Conclusions
Conclusions

Main results

- Both methods are beneficial in reverberant environments, also in conjunction with MC training, CMLLR, beamforming
- NMF approach outperforms the missing data methods
- Activation filtering degrades performance for clean speech

Future plans

- Missing data: improving the mask estimation
- NMF: convolutional NMF, activation matrix filtering
- Tackling both noise and reverberation with NMF
- Use of uncertainty information
References


Samples and sources

http://research.spa.aalto.fi/speech/robust/kallasjoki-reverb14/
Questions